Offline Reinforcement Learning with Implicit Q-Learning

Ilya Kostrikov, Ashvin Nair, Sergey Levine
University of California, Berkeley
International Conference on Learning Representations (ICLR 2022)

2026.01.08.
에이전트브레인스토밍 스터디
김동민

https://arxiv.org/abs/2110.06169

What's the problem of offline reinforcement learning?

$$L_{\mathrm{TD}}(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[\left(r(s,a) + \gamma \max_{a'} Q_{\hat{\theta}}(s',a') - Q_{\theta}(s,a)\right)^2\right]$$

- Offline RL이 어려운 이유
  - 고정된 데이터 (no exploration)
    - 정책 개선이 쉽지 않음 → unexplored action의 Q-value 평가 필요
    - → distributional shift & Q overestimation 문제 발생
- 기존 접근
  - Policy constraint
    - TD3+BC, AWAC
  - Q regularization
    - CQL
  - Single-step methods
    - DT (value iteration 아예 없음)
    - One-step RL (Brandfonbrener et al., 2021, Bellman backup을 한 번만 사용 behavior policy 평가)
- Key question
  - OOD action을 아예 평가하지 않으면서도(implicit) policy improvement가 가능할까?

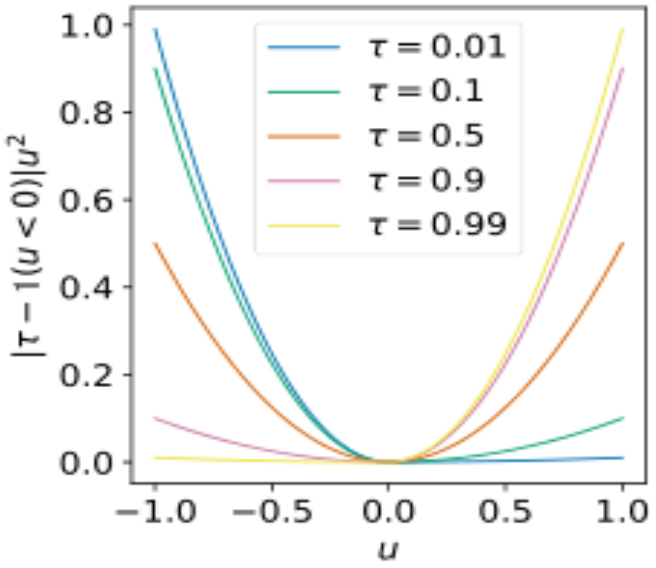
Key Insight of Implicit Q-Learning (IQL)

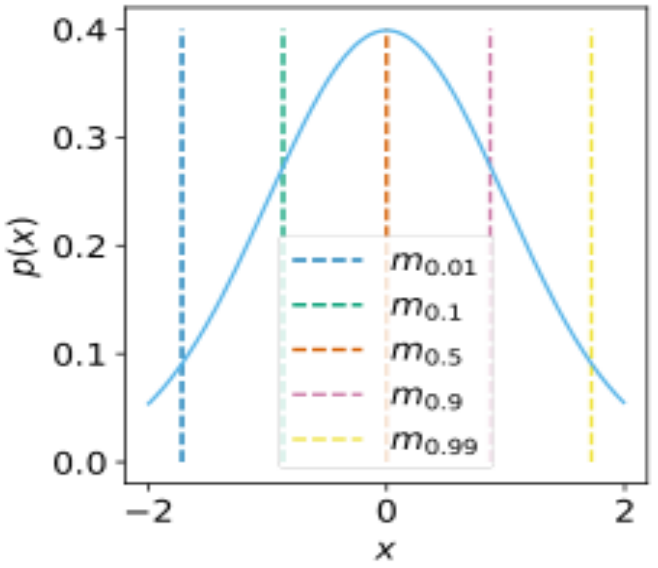
- Policy Improvement를 명시적으로 하지 않음
- 대신,
  - action에 대한 Q-value 분포의 **upper expectile**을 value로 사용
  - max operator를 **expectile regression**으로 근사

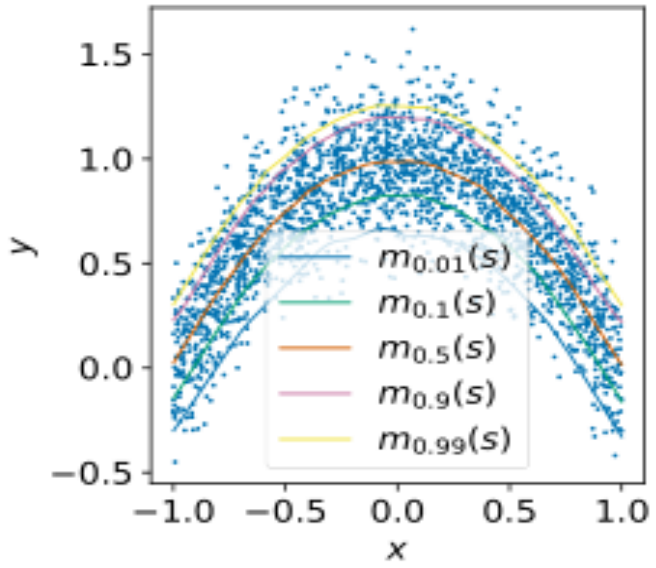
- Expectile Regression
  - Expectile: 비대칭 L2 loss 기반 통계량
  - $\tau = 0.5$ → 평균
  - $\tau \to 1$ → 상위 값 강조(≈ max)

$$L_\tau(u) = |\tau - \mathbf{1}_{u<0}| u^2$$

Max over actions ≒ high expectile of Q-values







IQL similar offline RL (1)

- First modification
  - SARSA-style objective
  - learn the value of the dataset policy (behavior policy) π_β

$$L_{\mathrm{TD}}(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[\left(r(s,a) + \gamma \max_{a'} Q_{\hat{\theta}}(s',a') - Q_\theta(s,a)\right)^2\right]$$

$$\Downarrow$$

$$L(\theta) = \mathbb{E}_{(s,a,s',a')\sim\mathcal{D}}\left[\left(r(s,a) + \gamma Q_{\hat{\theta}}(s',a') - Q_\theta(s,a)\right)^2\right]$$

  - never query values for out-of-sample actions
  - assuming unlimited capacity and no sampling error

$$Q_{\theta^*}(s,a) \approx r(s,a) + \gamma \mathbb{E}_{\substack{s'\sim p(\cdot|s,a) \\ a'\sim\pi_\beta(\cdot|s)}}\left[Q_{\hat{\theta}}(s',a')\right]$$

    - 기존 연구(Brandfonbrener et al., 2021; Peng et al., 2019)에서 활용한 objective
    - MuJoCo locomotion task에서는 잘 동작하지만 AntMaze에서는 잘 동작 안함
    - AntMaze와 같은 환경을 풀기 위해서는 multi-step dynamic programming을 수행해야 함

IQL similar offline RL (2)

- second modification
  - aim to estimate the maximum Q-value over actions that are in the support of the data distribution
    - support: 확률변수가 가질 수 있는 값의 범위, 이 경우에는 dataset으로 표현되는 액션값의 범위
  - without ever querying the learned Q-function on out-of-sample actions by utilizing expectile regression
  - modified objective

$$L_{\text{TD}}(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[\left(r(s,a) + \gamma \max_{a'} Q_{\hat{\theta}}(s',a') - Q_\theta(s,a)\right)^2\right]$$

$$\Downarrow$$

$$L_{\text{SARSA-like}}(\theta) = \mathbb{E}_{(s,a,s',a')\sim\mathcal{D}}\left[\left(r(s,a) + \gamma Q_{\hat{\theta}}(s',a') - Q_\theta(s,a)\right)^2\right]$$

$$\Downarrow$$

$$L_{\text{BCQ}}(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[\left(r(s,a) + \gamma \max_{\substack{a'\in\mathcal{A} \\ \text{s.t. } \pi_\beta(a'|s')>0}} Q_{\hat{\theta}}(s',a') - Q_\theta(s,a)\right)^2\right]$$

  - $\pi_\beta(a|s) > 0$: 상태 s에서 행동 a가 데이터셋에서 한 번이라도 등장했을 확률이 양수
    - 데이터셋의 support에 포함된 action만 고려
  - BCQ (Fujimoto et al., 2019), CQL (Kumar et al., 2020)

IQL (1)

- IQL's modification
  - use expectile regression that approximates the maximum
  - IQL's objective

$$L_{\text{TD}}(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[\left(r(s,a)+\gamma \max_{a'} Q_{\hat{\theta}}(s',a') - Q_\theta(s,a)\right)^2\right]$$

$$\Downarrow$$

$$L_{\text{SARSA-like}}(\theta) = \mathbb{E}_{(s,a,s',a')\sim\mathcal{D}}\left[\left(r(s,a)+\gamma Q_{\hat{\theta}}(s',a') - Q_\theta(s,a)\right)^2\right]$$

$$\Downarrow$$

$$L_{\text{BCQ}}(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[\left(r(s,a)+\gamma \max_{\substack{a'\in\mathcal{A} \\ \text{s.t. } \pi_\beta(a'|s')>0}} Q_{\hat{\theta}}(s',a') - Q_\theta(s,a)\right)^2\right]$$

$$\Downarrow$$

$$L_{\text{IQL}}(\theta) = \mathbb{E}_{(s,a,s',a')\sim\mathcal{D}}\left[L_2^\tau\left(r(s,a)+\gamma Q_{\hat{\theta}}(s',a') - Q_\theta(s,a)\right)\right]$$

IQL (2)

- IQL's further modification

$$L_{\text{IQL}}(\theta) = \mathbb{E}_{(s,a,s',a')\sim\mathcal{D}}\left[L_2^{\tau}\left(r(s,a)+\gamma Q_{\hat{\theta}}(s',a')-Q_{\theta}(s,a)\right)\right]$$

  - TD target에 transition stochasticity 포함
  - 어쩌다가 운좋은 transition이 max처럼 반영됨
  - → 행동에 대한 평균, 즉 $V(s)$를 활용
  - We resolve this by introducing a separate value function that approximates an expectile only with respect to the action distribution, leading to the following loss:

$$L_V(\psi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[L_2^{\tau}\left(Q_{\hat{\theta}}(s,a)-V_{\psi}(s)\right)\right]$$

  - We can then use this estimate to update the Q-functions with the MSE loss, which averages over the stochasticity from the transitions and avoids the "lucky" sample issue mentioned above:

$$L_Q(\theta) = \mathbb{E}_{(s,a,s')\sim\mathcal{D}}\left[\left(r(s,a)+\gamma V_{\psi}(s')-Q_{\theta}(s,a)\right)^2\right]$$

- 이 기법의 Optimality 증명은 Section 4.4에서 다루며 본 발표에서는 생략함
- 그런데 action-value function이 아니라 value function을 이용하기 때문에 policy가 명시적으로 드러나지 없음 → policy extraction 필요

IQL (3)

- Policy extraction
  - advantage-weighted regression (AWR) 기법 활용

$$L_\pi(\phi) = \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\exp\left(\beta\left(Q_{\hat{\theta}}(s,a) - V_\psi(s)\right)\right)\log \pi_\phi(a|s)\right]$$

  - β: inverse temperature
    - smaller → BC
    - larger → greedy Q-learning
- Algorithm summary

Algorithm 1 Implicit Q-learning

Initialize parameters ψ, θ, $\hat{\theta}$, ϕ.
TD learning (IQL):
for each gradient step **do**
$\psi \leftarrow \psi - \lambda_V \nabla_\psi L_V(\psi)$
$\theta \leftarrow \theta - \lambda_Q \nabla_\theta L_Q(\theta)$
$\hat{\theta} \leftarrow (1-\alpha)\hat{\theta} + \alpha\theta$
end for
Policy extraction (AWR):
for each gradient step **do**
$\phi \leftarrow \phi - \lambda_\pi \nabla_\phi L_\pi(\phi)$
end for

Evaluations (1)

- IQL is comparable to or better than others, especially with suboptimal datasets (antmaze)
  - near-optimal trajectory가 있는 데이터셋: MuJoCo locomotion, 일부 Kitchen / Adroit
  - near-optimal trajector가 없는 데이터셋: antmaze medium / large, antmaze diverse
- Computationally efficient: perform 1M updates on one GTX1080 GPU in less than 20 minutes
  - implemented in JAX

Dataset	BC	10%BC	DT	AWAC	Onestep RL	TD3+BC	CQL	IQL (Ours)
halfcheetah-medium-v2	42.6	42.5	42.6	43.5	**48.4**	**48.3**	44.0	**47.4**
hopper-medium-v2	52.9	56.9	**67.6**	57.0	59.6	59.3	58.5	**66.3**
walker2d-medium-v2	75.3	75.0	74.0	72.4	**81.8**	83.7	72.5	78.3
halfcheetah-medium-replay-v2	36.6	40.6	36.6	40.5	38.1	**44.6**	**45.5**	**44.2**
hopper-medium-replay-v2	18.1	75.9	82.7	37.2	**97.5**	60.9	**95.0**	**94.7**
walker2d-medium-replay-v2	26.0	62.5	66.6	27.0	49.5	**81.8**	77.2	73.9
halfcheetah-medium-expert-v2	55.2	**92.9**	86.8	42.8	**93.4**	**90.7**	**91.6**	86.7
hopper-medium-expert-v2	52.5	**110.9**	**107.6**	55.8	103.3	98.0	**105.4**	91.5
walker2d-medium-expert-v2	**107.5**	**109.0**	**108.1**	74.5	**113.0**	**110.1**	**108.8**	**109.6**
locomotion-v2 total	466.7	**666.2**	**672.6**	450.7	**684.6**	**677.4**	**698.5**	**692.4**
antmaze-umaze-v0	54.6	62.8	59.2	56.7	64.3	78.6	74.0	**87.5**
antmaze-umaze-diverse-v0	45.6	50.2	53.0	49.3	60.7	71.4	**84.0**	62.2
antmaze-medium-play-v0	0.0	5.4	0.0	0.0	0.3	10.6	61.2	**71.2**
antmaze-medium-diverse-v0	0.0	9.8	0.0	0.7	0.0	3.0	53.7	**70.0**
antmaze-large-play-v0	0.0	0.0	0.0	0.0	0.0	0.2	15.8	**39.6**
antmaze-large-diverse-v0	0.0	6.0	0.0	1.0	0.0	0.0	14.9	**47.5**
antmaze-v0 total	100.2	134.2	112.2	107.7	125.3	163.8	303.6	**378.0**
total	566.9	800.4	784.8	558.4	809.9	841.2	1002.1	**1070.4**
kitchen-v0 total	**154.5**	-	-	-	-	-	144.6	**159.8**
adroit-v0 total	104.5	-	-	-	-	-	93.6	**118.1**
total+kitchen+adroit	825.9	-	-	-	-	-	1240.3	**1348.3**
runtime	10m	10m	960m	20m	$\approx$ 20m*	20m	80m	20m

Evaluations (2)

- Online Fine-Tuning after Offline RL

Dataset	AWAC	CQL	IQL (Ours)
antmaze-umaze-v0	56.7 → 59.0	70.1 → **99.4**	**86.7** → **96.0**
antmaze-umaze-diverse-v0	49.3 → 49.0	31.1 → **99.4**	**75.0** → 84.0
antmaze-medium-play-v0	0.0 → 0.0	23.0 → 0.0	**72.0** → **95.0**
antmaze-medium-diverse-v0	0.7 → 0.3	23.0 → 32.3	**68.3** → **92.0**
antmaze-large-play-v0	0.0 → 0.0	1.0 → 0.0	**25.5** → **46.0**
antmaze-large-diverse-v0	1.0 → 0.0	1.0 → 0.0	**42.6** → **60.7**
antmaze-v0 total	107.7 → 108.3	151.5 → 231.1	**370.1** → **473.7**